

Abstract: Unlike the English language, languages such as German, Dutch, the Skandinavian languages or Greek form compounds not as multi-word expressions, but by combining the parts of the compound into a new word without any orthographical separation. This poses problems for a variety of tasks, such as Statistical Machine Translation or Information Retrieval. Most previous work on the subject of splitting compounds into their parts, or “decompounding” has focused on German. In this work, we create a new, simple, unsupervised system for automatic decompounding for three representative compounding languages: German, Swedish, and Hungarian. A multi-lingual evaluation corpus in the medical domain is created from the EMEA corpus, and annotated with regards to compounding. Finally, several variants of our system are evaluated and compared to previous work.